20

A Nonlinear Squeezing of the Continuous Wavelet Transform Based on Auditory Nerve Models

Ingrid Daubechies[1] and Stéphane Maes[2]

[1] *Program in Applied and Computational Mathematics and Department of Mathematics, Princeton University, Princeton, NJ*
[2] *IBM T. J. Watson Research Center, Human Language Technologies, Acoustic Processing Department, Yorktown Heights, NY*

20.1 Introduction

The approach presented in this chapter resulted from a concrete problem in speaker identification. Our goal was to incorporate the wavelet transform and auditory nerve-based models into a tool that could be used for speaker identification (among other applications), in the hope that the results would be more robust to noise than the standard methods.

This chapter is organized as follows. Sections 20.2 to 20.4 present background material, explaining, respectively, (1) how the (continuous) wavelet transform comes up "naturally" in our auditory system; (2) a heuristic approach (the ensemble interval histogram of O. Ghitza [1]) based on auditory nerve models, which eliminates much of the redundancy in the first-stage transform; and (3) the modulation model, valid for large portions of (voiced) speech, and which is used for speaker identification.[1] In Section 20.5 we put all this background material to use in our own synthesis, an approach that we call "squeezing" the wavelet transform; with an extra refinement this becomes "synchrosqueezing." The main idea is that the wavelet transform itself has "smeared" out different harmonic components, and that we need to "refocus" the resulting time-frequency or timescale picture. How this is done is explained in Section 20.5. Section 20.6 deals with various implementation issues, which are touched upon rather than explained in detail; for details, we refer to the various articles [2–6]. Finally, Section 20.7 shows some results: the "untreated" wavelet transform of a speech segment, its squeezed and synchrosqueezed versions, and the extraction of the parameters used for speaker identification. We conclude with some pointers to and comparisons with similar work in the literature, and with sketching possible future directions.

20.2 The Wavelet Transform as an Approach to Cochlear Filtering

When a sound wave hits our eardrum, the oscillations are transmitted to the basilar membrane in the cochlea. The cochlea is rolled up like a spiral; imagine unrolling it (and with it the basilar membrane), and putting an axis y onto it, so that points on the basilar membrane are labeled by their distance to one end. (For simplicity, we use a one-dimensional model, neglecting any influence of the transverse direction on the membrane, or its thickness.) If a pure tone, i.e., an excitation of the form $e^{i\omega t}$ (or its real part) hits the eardrum, then the response at the level of the basilar membrane, as observed experimentally or computed via detailed models, is in first approximation given by $e^{i\omega t}F_\omega(y)$ — a temporal oscillation with the same frequency as the input, but with an amplitude localized within a specific region in y by the function $F_\omega(y)$. In a first approximation, the dependence of F_ω on ω can be modeled by a logarithmic shift: $F_\omega(y) = F(y - \log\omega)$. (Strictly speaking, this model is only good for frequencies above say, 500 Hz; for low frequencies, the dependence of F_ω on ω is approximately linear.)

[1] Some of our descriptions of the auditory system may well look naïve and distorted to the more informed reader. They are in no way meant as an accurate description of what we realize is a very complex system. Rather, they are snapshots that motivated our mathematical construction further on, and they should be taken only as such.

The response to a more complicated $f(t)$ can then be computed as follows:

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{i\omega t}\,dw$$

$$\Longrightarrow \quad \text{response } B(t,y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{i\omega t}F(y-\log\omega)\,dw.$$

(Note that we are assuming linearity here — a superposition of inputs leading to the same superposition of the respective responses. This is again only a first approximation; richer and more realistic auditory models contain significant nonlinearities [7].) If we relabel the axis along the basilar membrane by defining $y := -\log a$ with $a > 0$ and $B'(t,a) = B(t, -\log a)$, and if we moreover define a function G by putting $F(x) =: \hat{G}(e^{-x})$, then the response can be rewritten as

$$B'(t,a) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t')e^{i\omega(t-t')}\hat{G}(a\omega)\,dt'\,dw \qquad (20.1)$$

$$= \int_{-\infty}^{\infty} f(t')\,\frac{1}{a}\,G\left(\frac{t-t'}{a}\right)\,dt'.$$

By taking $\psi(t) := G(-t)$, we find that $B'(t,a) = |a|^{-\frac{1}{2}}(W_\psi f)(a,t)$, where W_ψ is the continuous wavelet transform as defined by formula (1.23) in Chapter 1. In this sense, the cochlea can be seen as a "natural" wavelet transformer; all this is of course a direct consequence (and nothing but a reformulation) of the logarithmic dependence on ω of F_ω.

20.3 A Model for the Information Compression after the Cochlear Filters

The cochlear filtering, or the continuous wavelet transform that approximates it, transforms the one-dimensional signal $f(t)$ into a two-dimensional quantity. If we were to sample this two-dimensional transform like an image, then we would end up with an enormous number of data, far more than can in fact be handled by the auditory nerve. Some compression therefore has to take place immediately. The ensemble interval histogram (EIH) method of Oded Ghitza [1] gives such a compression, inspired by auditory nerve models. We describe it here in a nutshell, with its motivation.

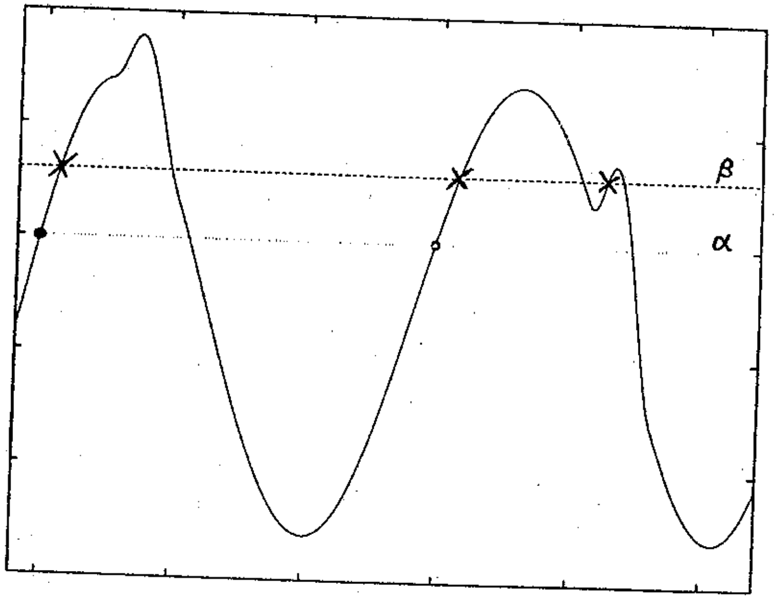

Figure 20.1
Displacement of the basilar membrane, at one fixed point y, as a function of time. The horizontal lines α and β represent the thresholds for bristles of different stiffness.

Near the basilar membrane, and over its whole length, one finds series of bristles of different stiffness. As the membrane moves near a particular bristle, it can, if the displacement is sufficiently large, "bend" the bristle. For different degrees of stiffness, this happens for different thresholds of displacement. Every time a bristle is bent, we think of this as an "event"; we also imagine that events only count when the bristle is bent away from its equilibrium position, not when it moves back. Figure 20.1 gives a schematic representation of what this means. The curve represents the movement of the membrane, as a function of time, at one particular location y.

The two horizontal lines, labeled by α and β, represent two different bristle thresholds, and the dots and crosses mark the corresponding "events" in the timespan represented in the figure. Replacing the information contained in all the curves (for different y) by only the coordinates (level, time, location) of these events would already be a formidable compression. Yet the EIH model reduces the information even more, by another transformation. Start by setting a certain resolution level ΔT, and a "window width" t_0. Then, for a given t, look back in time and count within the interval $[t - t_0, t]$, the number $N_{\alpha,y}(T)$ of successive events (for the bristle at position y and with stiffness α) that were spaced apart by an interval between T and $T + \Delta T$. Next, compute $S(t, T)$, the sum over all α and y of these $N_{\alpha,y}(T)$. This new representation $S(t, T)$ of the original signal is still two-dimensional, like the original cochlear or wavelet filtering output; it is, however, often sampled more coarsely than the continuous wavelet transform. More important, from our point of view, than the compression that this represents, is the very nonlinear and adaptive transformation represented by $S(t,T)$, which can again be viewed as a time-frequency representation (a second look at the construction of $S(t,T)$ shows that T^{-1} plays the role of an instantaneous frequency). O. Ghitza [1] compared the performance of EIH-based tools for several types of discrimination tests (such as word spotting) with the results obtained from LPC (linear predictive coding, a hidden Markov model for speech); for clean speech, LPC performed better, but the EIH-based schemes were, like the human auditory system itself, much more robust when the noise level was raised, and provided still useful results at noise levels where LPC could no longer be trusted. The nonlinear squeezing of the continuous wavelet transform that we describe in Section 20.5 is inspired by the EIH-construction.

20.4 The Modulation Model for Speech

The modulation model represents speech signals as a linear combination of amplitude and phase modulated components,

$$f(t) = \sum_{k=1}^{K} A_k(t) \cos[\theta_k(t)] + \eta(t),$$

where $A_k(t)$ is the instantaneous amplitude and

$$\omega_k(t) = \frac{d}{dt}\theta_k(t)$$

the instantaneous frequency of component (or formant) k; $\eta(t)$ takes into account the errors of modeling [8, 9]. In a slightly more sophisticated model, the components are viewed as "ribbons" in the time-frequency plane rather than "curves," and one also associates instantaneous bandwidths $\Delta\omega_k(t)$ to each component. The parameters $A_k(t)$, $\omega_k(t)$, and $\Delta\omega_k(t)$ are all assumed to vary in time (as the notation indicates), but we assume that this variation is slow when compared with the oscillation time of each component, measured by $[\omega_k(t)]^{-1}$. For large parts of speech, the modulation model is very satisfactory, and one can take $\eta(t) \simeq 0$; for other parts (e.g., fricative sounds) it is completely inadequate. The parameters $A_k(t)$, $\omega_k(t)$, and $\Delta\omega_k(t)$ (for those portions of speech where they are meaningful) can be used for speaker recognition. The basic idea is as follows. Imagine that the speech signal can be well represented by, say, $K = 8$ components. For each component, we have 3 parameters that vary in time. The signal can thus be viewed as a path in an $8 \times 3 = 24$-dimensional space. This path depends of course on both the speaker and the utterance. During certain portions (such as within one vowel), the 24 parameters remain in the same neighborhood, after which they make a rapid transition to another neighborhood, where they then dwell for a while, and so on. The order in which these "islands" appear depends on the utterance, but their location in our 24-dimensional space is believed to be independent of the utterance, and can be used to characterize the speaker. To use this for a speaker identification project, one must thus do two things: (1) extract the $A_k(t)$, $\omega_k(t)$, $\Delta\omega_k(t)$ (or a subset of these parameters) from the speech signal; and (2) process this information in a classification scheme in order to identify the speaker. When LPC methods are used for this purpose [10–12], one determines in fact only the $\omega_k(t)$ and $\Delta\omega_k(t)$, not the amplitudes $A_k(t)$. They are incorporated into one complex number,

$$z_k(t) = e^{i[\omega_k(t) + i\Delta\omega_k(t)]};$$

the $z_k(t)$ are the poles of the vocal tract transfer function

$$\mathcal{H}(z,t) = \sum_{k=1}^{K} \frac{1}{1 - z/z_k(t)}.$$

It is not always straightforward to label the $z_k(t)$ correctly with the LPC method, i.e., to decide which of the poles, determined separately, belongs to which component. To circumvent this, one works not with the $z_k(t)$ themselves, but with the so-called LPC-derived cepstrum,

$$c_n(t) = \frac{1}{n} \sum_{k=1}^{K} [z_k(t)]^n,$$

for which the exact attribution of the $z_k(t)$ does not matter; this formula is due to Schroeder [16]. This speaker identification program was developed at CAIP (Center for Aids to Industrial Productivity) at Rutgers University, by K. Assaleh, R. Mammone, and J. Flanagan [10–12]. Once the cepstrum is extracted, they use a neural network to do the classification and identification part. They fine-tuned it until it performed so well that it could perfectly distinguish identical twins, when starting from clean speech signals, thus outperforming most humans!

20.5 Squeezing the Continuous Wavelet Transform

Our goal is to use the continuous wavelet transform to extract reliably the different components of the modulation model (when it is applicable) and the parameters characterizing them. Our first problem is that the wavelet transform gives a somewhat "blurred" time-frequency picture. Let us take, for instance, a purely harmonic signal,

$$f(t) = A \cos \Omega t.$$

We compute its continuous wavelet transform $(W_\psi f)(a,b)$, using a wavelet ψ that is concentrated on the positive frequency axis (i.e., support $(\hat{\psi}) \subset [0,\infty)$, or $\hat{\psi}(\xi) = 0$ for $\xi < 0$; note that this means that ψ is complex):

$$\begin{aligned}(W_\psi f)(a,b) &= \int f(t)\, \frac{1}{\sqrt{a}}\, \overline{\psi\left(\frac{t-b}{a}\right)}\, dt \qquad (20.2)\\ &= \frac{1}{2\pi} \int \hat{f}(\xi)\, \sqrt{a}\, \overline{\hat{\psi}(a\xi)}\, e^{ib\xi}\, d\xi \\ &= \frac{1}{2\pi} \int \frac{A}{2}\, [\delta(\xi - \Omega) + \delta(\xi + \Omega)]\, \sqrt{a}\, \overline{\hat{\psi}(a\xi)}\, e^{ib\xi}\, d\xi \\ &= \frac{A}{4\pi}\, \sqrt{a}\, \overline{\hat{\psi}(a\Omega)}\, e^{ib\Omega}.\end{aligned}$$

If $\hat{\psi}(\xi)$ is concentrated around $\xi = 1$, then $(W_\psi f)(a,b)$ will be concentrated around $a = \Omega^{-1}$, as expected. But it will be spread out over a region around this value (see Figure 20.2), and not give a sharp picture of what was a signal very sharply localized in frequency.

In order to remedy this blurring, the "Marseilles group" developed the so-called "ridge and skeleton" method [13]. In this method, special curves (the ridges) are singled out in the (a,b)-plane, depending on the wavelet transform $(W_\psi f)(a,b)$ itself (for each b, one finds the values of a where the oscillatory integrand in $(W_\psi f)(a,b)$ has "stationary phase"; for the signals considered here, this amounts to ∂_b[phase of $(W_\psi f)(a,b)] = \omega_0/a$, where ω_0 is the center frequency for ψ). From the restriction of $W_\psi f$ to these ridges (the "skeleton" of the wavelet transform), one can then read off the important parameters, such as the instantaneous frequency. This method has been used with great success for various applications, such as reliably

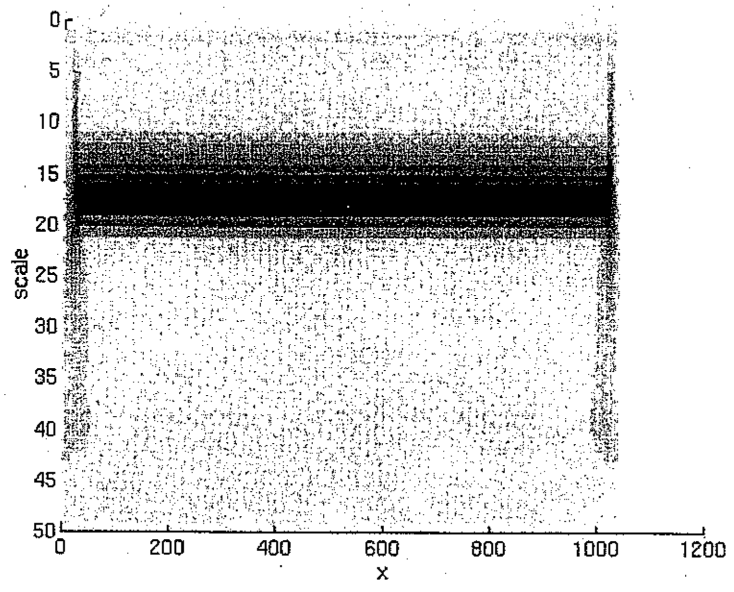

Figure 20.2
Absolute value $|W_\psi f(a,b)|$ of the wavelet transform of a pure tone f.

identifying and extracting spectral lines of widely different strengths [13]. In our speech signals, we have many components, some of which can remain very close for a while, to separate later again; components can also die or new components can suddenly appear out of nowhere. For these signals, the ridge and skeleton method does not perform as well. For this reason, we developed a different approach, where we try to squeeze back the defocused information in order to gain a sharper picture; in so doing, we try to use the whole wavelet transform instead of concentrating on special curves.

Let us look back at the wavelet transform (20.2) of a pure tone. Although it is spread out over a region in the a-variable around $a = \Omega^{-1}$, the b-dependence still shows the original harmonic oscillations with the correct frequency, regardless of the value of a. This suggests that we compute, for any (a,b), the instantaneous frequency $\omega(a,b)$ by

$$\omega(a,b) = -i[W_\psi f(a,b)]^{-1}\frac{\partial}{\partial b}W_\psi f(a,b),$$

and that we transfer the information from the (a,b)-plane to a (b,ω)-plane, by taking for instance,

$$S_\psi f(b,\omega_\ell) = \sum_{a_k \text{ such that } |\omega(a_k,b)-\omega_\ell| \le \Delta\omega/2} |W_\psi f(a_k,b)|. \qquad (20.3)$$

We have assumed here that both the old a-variable and the new ω-variable have been discretized. (A continuous formulation would be to introduce, for every b, a measure $d\mu_b$ in the ω-variable, which assigns to Borel sets A the measure

$$\mu_b(A) = \int |W_\psi f(a,b)| \chi_A(\omega(a,b))\, da,$$

where χ_A is the indicator function of A, $\chi_A(u) = 1$ if $u \in A$, $\chi_A(u) = 0$ if $u \notin A$.) This has *exactly* the same flavor as the EIH transform described in Section 20.3: we transform to a different time-frequency plane by reassigning contributions with the same instantaneous frequency to the same bin, and we give a larger weight to components with large amplitude $|W_\psi f|$ (just as components with large amplitude in the EIH would give rise to several level crossings and would therefore contribute more). Our S_ψ is also close to the SBS (in-synchrony bands spectrum, a precursor of the EIH) [14] or to the IFD (instantaneous frequency distribution) [15]. For good measure, one can also sum the $|a_k|^{-\alpha}|W_\psi f(a_k,b)|$ rather than the $|W_\psi f(a_k,b)|$, thus renormalizing the fine-scale regions where often $|W_\psi f(a,b)|$ is much smaller.

When this squeezing operation is performed on the wavelet transform of a pure tone, we find a single horizontal line in the (b,ω)-plane, at $\omega = \Omega$, as expected.

We can, however, refine the operation even further, and define a particular type of squeezing, which we call *synchrosqueezing*, that still allows for reconstruction, even after the (highly nonlinear!) transformation. To see this, we first have to observe that the reconstruction formula of f from $W_\psi f$, given by formula (1.25) in Chapter 1, is not the only one. We also have, again assuming support $\hat{\psi} \subset [0,\infty)$,

$$\begin{aligned} \int_0^\infty W_\psi f(a,b)\, a^{-3/2}\, da &= \int\int \hat{f}(\xi)\, e^{ib\xi}\, \overline{\hat{\psi}(a\xi)}\, a^{-1}\, da\, d\xi \qquad (20.4) \\ &= \left[\int_0^\infty \hat{\psi}(\xi)\, \frac{d\xi}{\xi}\right] \cdot \int \hat{f}(\xi) e^{ib\xi}\, d\xi \\ &= \left[2\pi \int_0^\infty \hat{\psi}(\xi)\, \frac{d\xi}{\xi}\right] f(b). \end{aligned}$$

This suggests that we define

$$(\mathcal{S}_\psi f)(b,\omega_\ell) = \sum_{a_k \text{ such that } |\omega(a_k,b) - \omega_\ell| \le \Delta\omega/2} W_\psi f(a_k,b) a_k^{-3/2} \qquad (20.5)$$

(without absolute values!); with ω_ℓ spaced apart by $\Delta\omega$, we then still have

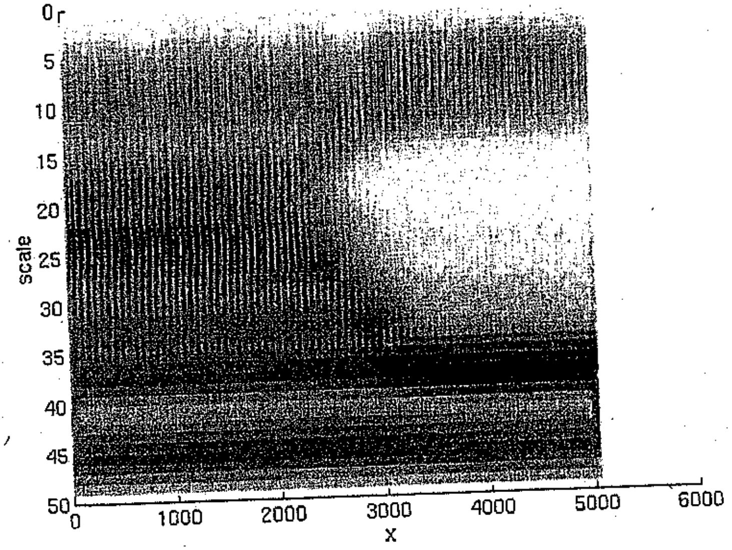

Figure 20.3
Absolute value $|W_\psi f(a,b)|$ of the sound /a-a–i-i/. A colored noise is present with SNR = 15 dB. The horizontal axis is sampled at 8 kHz. The vertical axis represents different subbands (five octaves, split into eight equally spaced suboctaves); low indices are associated to high frequencies.

(in the assumption that the discretizations are sufficiently fine to be good approximations to integrals)

$$\sum_{\ell}(\mathcal{S}_\psi f)(\omega_\ell, b) = C_\psi^{\#} f(b). \tag{20.6}$$

Having the exact reconstruction (20.6) will be useful to us later on (see the end of this section); note that such an exact reconstruction is not available for the EIH, SBS, or IFD. There is an added bonus to *synchro*squeezing. The process of reassigning components from the (a,b)-plane to the (b,ω)-plane is not perfect, especially when noise is present, and occasionally parts of components that are truly different get assigned to the same ω_ℓ-bin. When this happens, the two pieces from different components are often out of phase with each other, and cancellation takes place in the computation of $\mathcal{S}_\psi$ (but not in S_ψ!). Figures 20.3 and 20.4 show the unprocessed wavelet transform and the synchrosqueezed wavelet transform, respectively, of the speech signal consisting of the two vowels /a-a–i-i/; clearly, the different components can be distinguished much more clearly after the (syn-

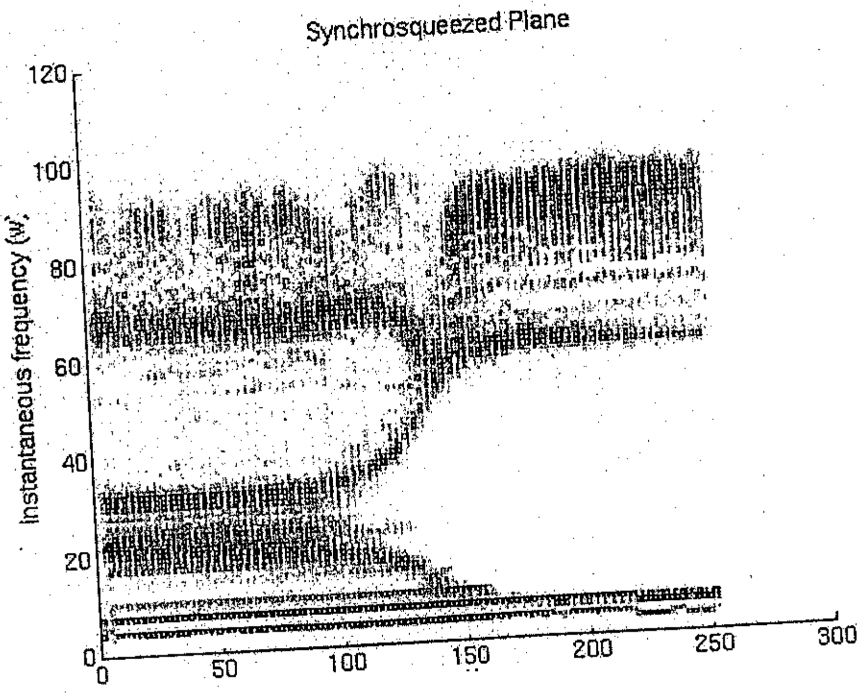

Figure 20.4
Synchrosqueezed representation of /a-a-i-i/ (same signal, same noise level as in Figure 20.3). The components can be distinguished much more clearly than in Figure 20.3. (Note that because the scale a corresponds to ω^{-1}, there is also a distortion of the vertical axis when compared to Figure 20.3.)

chro)squeezing. The extra focusing of the synchrosqueezing over squeezing can be seen in an example in Section 20.7.

One remark is in order here. Both the squeezing and synchrosqueezing operations can be defined with any arbitrary reassigning rule — it does not have to be governed by the instantaneous frequency. In particular, the reconstruction property from $S_\psi f$ does not depend on the physical interpretation of the reassignment rule. This means that we should not worry about the parts of f where the modulation model does not apply — true, the reassignment will not be as meaningful, because instantaneous frequency does not make much sense there, but we still haven't "hurt" the information that was there. In fact, as the synchrosqueezed representation of "august" in Figure 20.5 shows, the "s" part is still nicely localized in the upper frequencies, where it belongs, so in practice we don't seem to displace such nonmodulated parts in the time-frequency plane. Of course, the refocusing that we see in the squeezed and synchrosqueezed transform *does* depend on the physical interpretation — an arbitrary reassignment rule would give a messy picture.

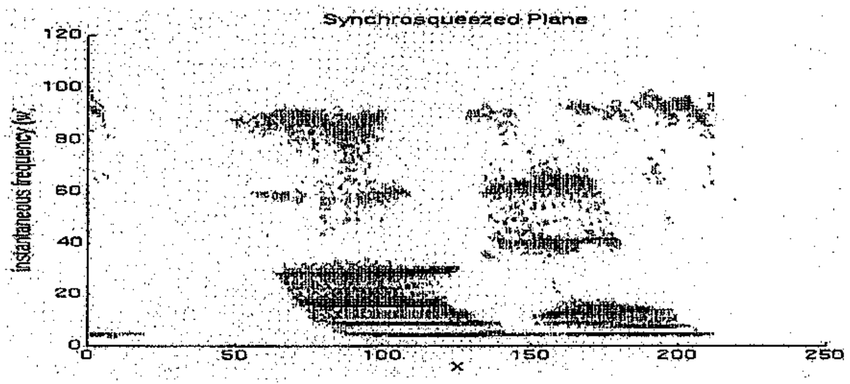

Figure 20.5
Synchrosqueezed representation of /ow-g-λ-s-t/. A colored noise is present with SNR of 15 dB. The "s" part is the cloud in the upper right corner.

After synchrosqueezing, the components are well-separated and can be identified. From the synchrosqueezed representation, we can determine the central frequencies $\omega_k(t)$ and the bandwidths $\Delta\omega_k(t)$. How can we find the $A_k(t)$? Remember our exact reconstruction formula (20.6)! If a post-processing step separates the different components in the synchrosqueezed plane, then we can carve out the component under consideration in the synchrosqueezed plane, delete all the rest, and reconstruct from only this component; this is called the selective fusion algorithm [2, 4]. The direct summation method (20.6) provides fast and relatively accurate results; a slightly slower but even more accurate method uses double integrals (see [2, 4]). This is carried out for speech signals, within the modulation model framework, in [2, 5]. From every reconstructed single component, we can then determine $A_k(t)$, $\theta_k(0)$ so that $A_k(t)\cos(\theta_k(t))$ fits this reconstructed component, within the constraint $\frac{d}{dt}\theta_k(t) = \omega_k(t)$.

This finishes our program of extracting the modulation model parameters from an EIH-analog based on the wavelet transform. After a (very summary) discussion of some implementation issues, we shall return to results in Section 20.7.

20.6 Short Discussion of Some Implementation Issues

First of all, the whole construction is based on a continuous wavelet transform. In practice, this is of course a discrete but very redundant transform, heavily oversampled both in time and in scale. In order to be practical, we need a fast implementation scheme. This was achieved by borrowing a leaf from (nonredundant) wavelet bases, i.e., by using subband filtering schemes. For a given profile $\hat{\psi}(\xi)$ (close to that of a Morlet wavelet), we identified a function $\hat{\phi}$ and trigonometric polynomials $\hat{h}, \hat{g}_\ell$; $\ell = 1, \ldots, L$, so that

$$\hat{\psi}(2^{(\ell-1)/L}\omega) \simeq \hat{g}_\ell(\omega)\hat{\phi}(\omega)$$

$$\hat{\phi}(2\omega) \simeq \hat{h}(\omega)\hat{\phi}(\omega).$$

This means that the Fourier coefficients of $\hat{h}, \hat{g}_\ell$ can be used for an iterated FIR filtering scheme that gives the redundant wavelet transform in linear time. For details on the algorithm and on the construction of the filters, see [2, 6], or Chapter 2, Section 2.5 in this book.

Next we note that the squeezing and synchrosqueezing operations entailed first the determination of the instantaneous frequency $\omega(a, b)$. This was done by a logarithmic differentiation of $W_\psi f(a, b)$. This is of course very unstable when $|W_\psi f(a, b)|$ is small; note however that these regions will contribute very little to either $S_\psi f$ or $\mathcal{S}_\psi f$ (defined by (20.3) and (20.5), respectively), so that we can safely avoid this problem by putting a lower threshold on $|W_\psi f(a, b)|$. On the other hand, differentiation itself is also a tricky business when the data are noisy; in practice, a standard numerical difference operator was used, involving a weighted differencing operator, spread out over a neighborhood of samples. Again, details can be found in [2, 4].

In the previous section, we also glossed over the extraction of the $\omega_k(t)$, $\Delta\omega_k(t)$ from the synchrosqueezed picture. In fact, although we can often clearly see the different components with our eyes, extracting them and their parameters automatically is a different matter. For instance, in "How are you?", an example shown in Section 20.7, the components are much weaker in some spots than in others, yet we want our "extractor" to bridge those weak gaps. The approach we use, developed with Trevor Hastie [18], views $|\mathcal{S}_\psi f(b, \omega)|$ as a probability distribution in ω, for every value of b, which can be modeled as a mixture of Gaussians, and which evolves as b changes; moreover, we impose that the centers of the Gaussians follow paths given by splines (cubic or linear). We also allow components to die or to be born. In order to find an evolution law that fits the given $|\mathcal{S}_\psi f(b, \omega)|$, a few steps of an iterative scheme suffice; for details, see [2, 18]. The resulting centers of the Gaussians in the mixture give us the frequencies $\omega_k(t)$, their widths give us the $\Delta\omega_k(t)$.

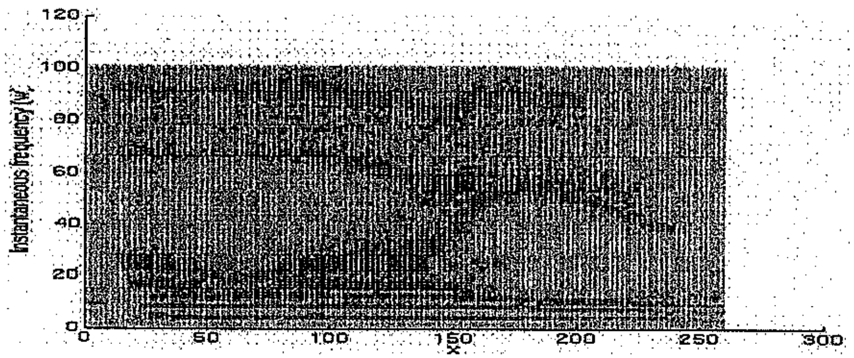

Figure 20.6
Squeezed plane representation for /h-δ-w-a-r-j-u?/. A colored noise is present with SNR = 15 dB.

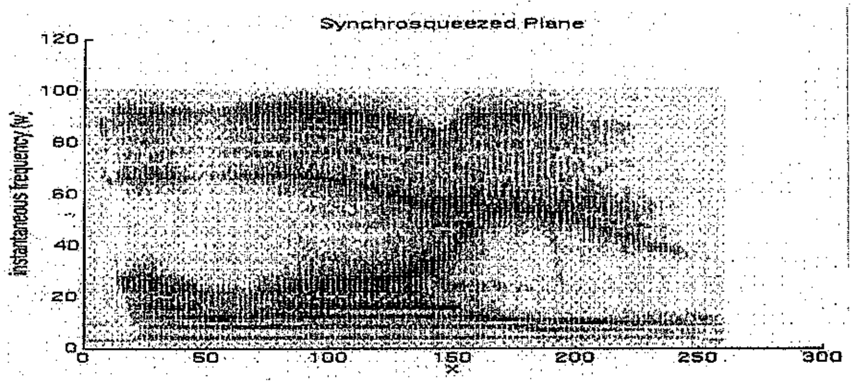

Figure 20.7
Synchrosqueezed plane representation for /h-δ-w-a-r-j-u?/. A colored noise is present with SNR = 15 dB.

20.7 Results on Speech Signals

We start by illustrating the enhanced focusing of the synchrosqueezed representation when compared to the squeezed representation of a different example, namely, the utterance, "How are you?" or /h-δ-w-a-r-j-u?/; see Figures 20.6 and 20.7.

Figure 20.8 shows the curves for the corresponding extracted central frequencies $\omega_k(t)$. In this case, the original signal was somewhat noisy; the (pink) noise had an SNR of about 15 dB.

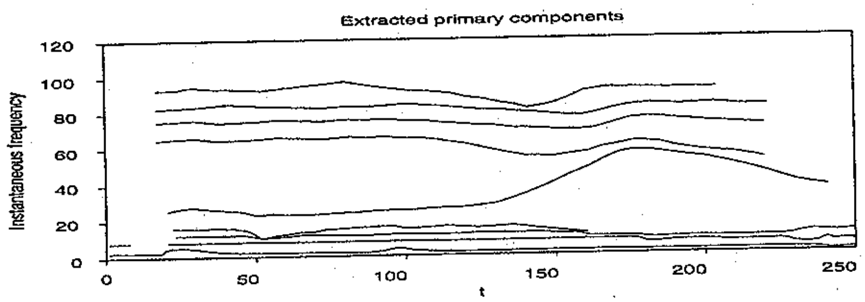

Figure 20.8
Curves for the central frequencies $\omega_k(t)$ for /h-δ-w-a-r-j-u?/. A colored noise is present with SNR = 15 dB.

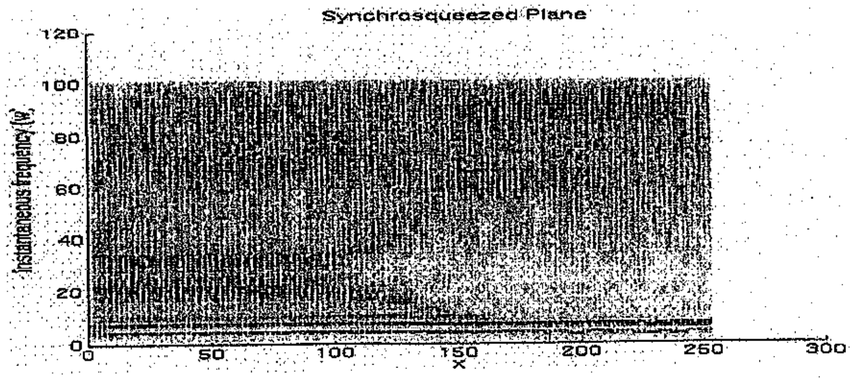

Figure 20.9
Synchrosqueezed plane representation for /$\cdots$-a-a–i-i-$\cdots$/. A colored noise is present with SNR = 15 dB. An additional white noise is added with SNR = 11 dB.

Next, we illustrate the robustness of our analysis under higher noise levels. We return to the signal /a-a–i-i/, this time with an additional white noise with SNR of 11 dB. Figure 20.9 shows the synchrosqueezed representation of this noisier signal; although the representation is noisier as well, the different components can still be identified clearly, and they haven't moved. This is borne out by a comparison of the extracted central frequency curves. Figure 20.10 shows the extracted frequency curves for the slightly noisy original of Figure 20.4. Figure 20.11 shows the extracted frequency curves for the much noisier version given in Figure 20.9.

Finally, we also show results of a first test of the use of the synchrosqueezed representation for speaker identification. For this first test, we did not use the full strength of the representation, and we did not develop our own classification either. Instead, we took our $\omega_k(t)$, $\Delta\omega_k(t)$ values,

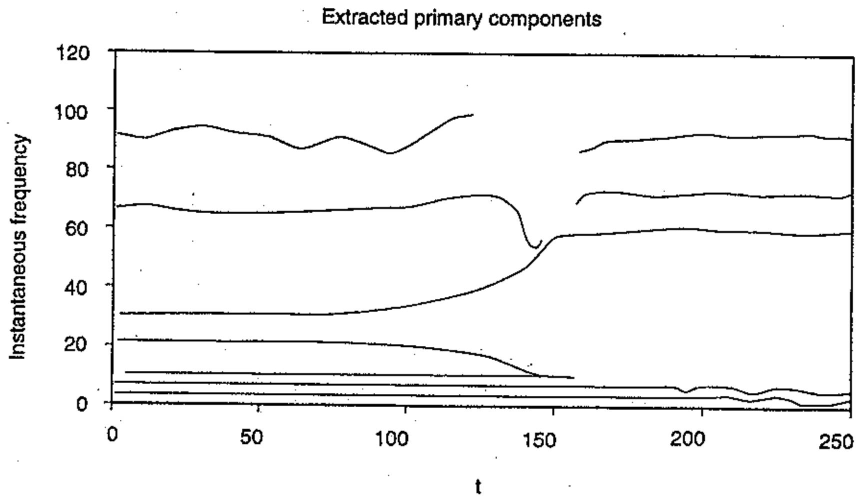

Figure 20.10
Curves for the central frequencies $\omega_k(t)$ for /a-a–i-i/, extracted from Figure 20.4.

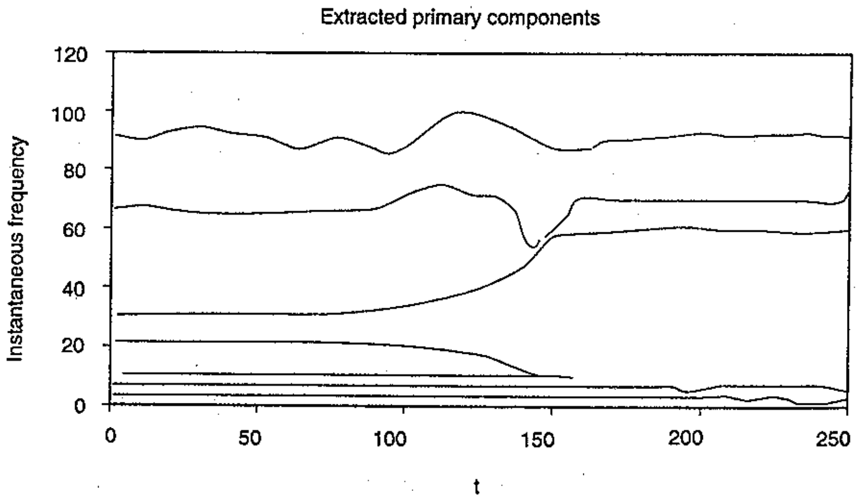

Figure 20.11
Curves for the central frequencies $\omega_k(t)$ for /a-a–i-i/ with additional white noise; see Figure 20.9.

and constructed an analog to the LPC-derived cepstrum by defining

$$z_k^w(t) = \exp[i\omega_k(t) - \Delta\omega_k(t)]$$

$$c_n^w(t) = \frac{1}{n}\sum_{k=1}^{K}[z_k^w(t)]^n\,;$$

we called this the "wastrum." We then used the wastrum as input for the classification scheme that had been developed at CAIP. For the experiment we performed, the input data come from the narrowband part of the *KING* database, released by ITT Aerospace/Communications Division, in April 1992. It is a telephone network database built with 52 American speakers, among whom the first 26 speakers are from the San Diego region. For each speaker, ten sessions have been recorded. The first five sessions were recorded at intervals of 1 week. Each session is narrowband, with the bandwidth of a telephone channel. Each session consists of roughly 50 to 75 seconds of conversational speech which contains roughly 40% of silences. The sessions are recorded from the interlocutor's side. The first five sessions are within the Great Divide, which means on the West Coast. The SNR is about 15 dB to 20 dB. This noise is introduced by the phone network. The five remaining sessions are recorded across the Great Divide at intervals of 1 month and they are much noisier. These last sessions were not used in this experiment. The signal is sampled at 8 kHz and quantized over 12 bits.

For the experiment, the first session of the first 26 speakers is used for training and the following four within divide sessions are used for testing.

The classifier is a vector quantizer. Decisions are made on the basis of the cumulated distances obtained in each frame relative to the codebooks associated to the different speakers.

Table 20.1 summarizes the results in closed-set speaker identification obtained with the LPC-derived cepstrum and the wastrum. The long-term mean is removed from the features, in agreement with [10]. The silence frames are removed on the basis of energy thresholds for the primary components. The same frames are removed for the LPC approach, in order to compare exactly the same utterances.

The performances of the wastrum are comparable to the LPC-derived cepstrum for the relatively clean speech, which is reassuring: we aim to extract the same cepstral-like information, albeit with very different methods, and so we expect similar performance! The wastrum method is, however, more robust,to noise when the noise can not be considered as negligible, since we get a lower error rate even though the noise level is significantly higher (12 dB versus 15 dB).

Table 20.1
Summary of the results obtained on *KING* database, within the Great Divide, 26 speakers, first section used for training, four other sessions used for testing. Long-term mean removal is used.

Method	additional SNR	error rate
LPC-derived cepstrum	none	~ 0.22
wastrum	none	0.23
LPC-derived cepstrum	15 dB	0.33
wastrum	12 dB	0.3

Note that we are comparing here a suboptimal version of our approach (the $A_k(t)$ are not taken into account, and the $\omega_k(t)$, $\Delta\omega_k(t)$ are transformed into the wastrum, that is then put through a classification scheme not specially tailored to our different approach) with a very much optimized version of the LPC-based method. Yet even so, the wastrum method leads to fewer errors for noisy speech than the LPC-derived cepstrum. This indicates that we have indeed inherited (some of) the robustness that characterizes true auditory systems.

The following is a short list of promising future directions to be explored: include the amplitude information $A_k(t)$ (obtained by selective fusion [5]) as well; develop a more direct classification scheme, without the detour of the wastrum, and maybe even directly from the synchrosqueezed plane, without extraction of the parameters first; and finally, use of this approach for other tasks in speech analysis.

There is some similarity between our squeezing and synchrosqueezing methods and a technique of "reassignment" developed by Auger and Flandrin [17], with the same goal of "refocusing" in the time-frequency plane; we first heard of their method after the work described here was completed. Auger and Flandrin typically work with Wigner-Ville or similar time-frequency distributions, and their reassignment method is not limited to one direction only (we don't change the b variable in our scheme); on the other hand, their scheme is not linked to an exact reconstruction formula such as our (20.6).

20.8 Acknowledgments

Both authors would like to thank CAIP (Rutgers University) and especially Prof. R. Mammone and his group for their hospitality and advice. Ingrid Daubechies also wants to thank NSF (grant DMS-9401785) and AFOSR (grant F49620-95-1-0290) for partial support.

References

[1] O. Ghitza. Auditory models and human performances in tasks related to speech coding and speech recognition. *IEEE Trans. Speech Audio Proc.*, 2(1):115–132, 1994; see also O. Ghitza. Advances in speech signal processing, in *Advances in speech signal processing*, S. Furui and M. Sondhi, editors. Marcel Dekker, New York, NY 1991.

[2] S. Maes. The wavelet transform in signal processing, with application to the extraction of the speech modulation model features. Ph.D. thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1994.

[3] S. Maes. The synchrosqueezed representation yields a new reading of the wavelet transform. In *Proc. SPIE 1995 on OE/Aerospace Sensing and Dual Use Photonics – Wavelet Applications for Dual Use – Session on Acoustic and Signal Processing, Wavelet Applications II*, Vol. 2491, H. H. Szu, editor, Orlando, FL, April 1995. Part I, 532–559

[4] S. Maes. The wavelet-derived synchrosqueezed plane representation yields a new time-frequency analysis of 1-D signals, with application to speech. *Preprint submitted to IEEE Trans. Speech and Audio Processing.*

[5] S. Maes. The wavelet-derived synchrosqueezed plane representation yields new front-ends for automated speech recognition. *Preprint submitted to IEEE Trans. Speech and Audio Processing.*

[6] S. Maes. Fast quasi-continuous wavelet algorithms for analysis and synthesis of 1-D signals. *Preprint submitted to SIAM J. Appl. Math.*

[7] J. B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, 2(1):3–29, 1985.

[8] C. D'Alessandro. Time-frequency speech transformation based on an elementary waveform representation. *Speech Commun.*, 9:419–431, 1990.

[9] J. S. Liénard. Speech analysis and reconstruction using short-time, elementary, waveforms. In *IEEE Proc. ICASSP*, Dallas, TX, 1987, 948–951

[10] K. Assaleh. *Robust features for speaker identification.* Ph.D. thesis, CAIP Center – Rutgers University, The State University of New Jersey, New Brunswick, NJ, 1993.

[11] K. Assaleh, R. J. Mammone, and J. L. Flanagan. Speech recognition using the modulation model. In *IEEE Proc. ICASSP*, Vol. 2, 1993, 664–667

[12] K. T. Assaleh and R. J. Mammone. New LP-derived features for speaker identification. *IEEE Trans. Speech Audio Proc.*, 2(4):630–638, 1994.

[13] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, Ph. Tchamitchian, and B. Torrésani. Asymptotic wavelet and Gabor analysis: extraction of instantaneous frequencies. *IEEE Trans. Inf. Theory*, 38(2 Part II):644–664, 1992.

[14] O. Ghitza. Auditory nerve representation criteria for speech analysis/synthesis. *IEEE Trans. ASSP*, 6(35):736–740, 1987.

[15] D. H. Friedman. Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech. In *IEEE Proc. ICASSP*, 1985, 1121–1124.

[16] M. Schroeder. Direct (non-recursive) relations between cepstrum and predictor coefficients. *IEEE Trans. ASSP*, 29:297–301, 1981.

[17] F. Auger and P. Flandrin. Improving the readablilty of time-scale representations by the reassignment method. *IEEE Trans. Signal Process.*, 43(5):1068–1089, 1995.

[18] T. Hastie and S. Maes. The maximum-likelihood-estimation-based living cubic spline extractor and its application to saliency grouping in the time-frequency plane. Preprint.